BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

As rescanning documents *will not* correct images, please do not report the images to the Image Problem Mailbox.

02.06.00
JP00/3623

4

日 本 国 特 許 庁
PATENT OFFICE
JAPANESE GOVERNMENT

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日
Date of Application: １９９９年 ６月１７日

27 JUL 2000

出 願 番 号
Application Number: 平成１１年特許願第１７１７２３号

出 願 人
Applicant (s): セイコーエプソン株式会社

PRIORITY DOCUMENT
SUBMITTED OR TRANSMITTED IN COMPLIANCE WITH RULE 17.1(a) OR (b)

２０００年 ６月２９日

特 許 庁 長 官
Commissioner,
Patent Office 近 藤 隆 彦

出証番号 出証特２０００－３０５２０４３

【書類名】　特許願

【整理番号】　J0074154

【提出日】　平成11年 6月17日

【あて先】　特許庁長官　殿

【国際特許分類】　G06F 7/24

【発明の名称】　情報分類方法及び情報分類装置並びに情報分類処理プログラムを記録した記録媒体

【請求項の数】　21

【発明者】

【住所又は居所】　長野県諏訪市大和３丁目３番５号　セイコーエプソン株式会社内

【氏名】　長石　道博

【特許出願人】

【識別番号】　000002369

【氏名又は名称】　セイコーエプソン株式会社

【代表者】　安川　英昭

【代理人】

【識別番号】　100093388

【弁理士】

【氏名又は名称】　鈴木　喜三郎

【連絡先】　０２６６－５２－３１３９

【選任した代理人】

【識別番号】　100095728

【弁理士】

【氏名又は名称】　上柳　雅誉

【選任した代理人】

【識別番号】　100107261

【弁理士】

【氏名又は名称】　須澤　修

【手数料の表示】

【予納台帳番号】　013044

【納付金額】　21,000円

【提出物件の目録】

【物件名】　明細書　1

【物件名】　図面　1

【物件名】　要約書　1

【包括委任状番号】　9711684

【プルーフの要否】　要

【書類名】　明細書

【発明の名称】　情報分類方法及び情報分類装置並びに情報分類処理プログラムを記録した記録媒体

【特許請求の範囲】

【請求項１】　検索サービスにユーザからの検索要求が与えられることによって検索された複数の文書に対し、それぞれの文書の共通性に基づいてクラスタリング処理し、それによって得られたクラスタリング結果に対し、検索されたそれぞれの文書対応に付された検索要求との適合性を示す値（スコアという）を用いて、前記クラスタリング処理によって得られたそれぞれのクラスタの順位を再構成し、そのクラスタ順位が再構成されたクラスタリング結果を出力することを特徴とする情報分類方法。

【請求項２】　前記それぞれのクラスタの順位を再構成する処理は、それぞれのクラスタに含まれるそれぞれの文書対応に付されたスコアの平均値をそれぞれのクラスタごとに求め、クラスタごとの平均値をそれぞれのクラスタのクラスタスコアとし、そのクラスタスコアによって、クラスタの順位を再構成することを特徴とする請求項１記載の情報分類方法。

【請求項３】　前記それぞれのクラスタの順位を再構成する処理は、それぞれのクラスタに含まれるそれぞれの文書対応に付されたスコアの最大値をそれぞれのクラスタごとに得て、そのクラスタごとのスコアの最大値をそれぞれのクラスタスコアとし、そのクラスタスコアによって、クラスタの順位を再構成することを特徴とする請求項１記載の情報分類方法。

【請求項４】　前記それぞれのクラスタの順位を再構成する処理は、それぞれのクラスタに含まれるそれぞれの文書がそれぞれの文書対応に付されたスコアの大きい順に並べられている場合、その中央または中央付近に位置するスコアをそれぞれのクラスタごとに得て、そのクラスタごとの中央または中央付近に位置するスコアをそれぞれのクラスタスコアとし、そのクラスタスコアによって、クラスタの順位を再構成することを特徴とする請求項１記載の情報分類方法。

【請求項５】　前記クラスタリング処理を複数の検索サービスによって得られた検索結果に対応して行うことを可能とする場合、前記クラスタの順位を再構

成するためのクラスタスコアを求める処理は、複数の検索サービスそれぞれに対応して行うことを特徴とする請求項２から４のいずれか１項に記載の情報分類方法。

【請求項６】　前記クラスタリング処理は、それぞれの文書のタイトルを検出し、そのタイトルに含まれる特徴的な用語を特徴要素として抽出し、抽出された特徴要素に基づいて行うことを特徴とする請求項１から５のいずれか１項に記載の情報分類方法。

【請求項７】　前記クラスタ順位が再構成されたクラスタリング結果の出力の仕方は、クラスタスコアの高いクラスタ順に表示し、クラスタスコアが同じであるクラスタが存在する場合には、クラスタ内の文書数の多いクラスタを高順位とすることを特徴とする請求項１から６のいずれか１項に記載の情報分類方法。

【請求項８】　検索サービスにユーザからの検索要求が与えられることによって検索された複数の文書に対し、それぞれの文書の共通性に基づいてクラスタリング処理するクラスタリングモジュールと、

このクラスタリングモジュールによって得られたクラスタリング結果に対し、検索されたそれぞれの文書対応に付された検索要求との適合性を示す値（スコアという）を用いて、前記クラスタリング処理によって得られたそれぞれのクラスタの順位を再構成し、そのクラスタ順位が再構成されたクラスタリング結果を出力するクラスタ順位設定モジュールと、

を有することを特徴とする情報分類装置。

【請求項９】　前記クラスタ順位設定モジュールが行うそれぞれのクラスタの順位を再構成する処理は、それぞれのクラスタに含まれるそれぞれの文書対応に付されたスコアの平均値をそれぞれのクラスタごとに求め、クラスタごとの平均値をそれぞれのクラスタのクラスタスコアとし、そのクラスタスコアによって、クラスタの順位を再構成することを特徴とする請求項８記載の情報分類装置。

【請求項１０】　前記クラスタ順位設定モジュールが行うそれぞれのクラスタの順位を再構成する処理は、それぞれのクラスタに含まれるそれぞれの文書対応に付されたスコアの最大値をそれぞれのクラスタごとに得て、そのクラスタごとのスコアの最大値をそれぞれのクラスタスコアとし、そのクラスタスコアによ

って、クラスタの順位を再構成することを特徴とする請求項８記載の情報分類装置。

【請求項１１】　前記クラスタ順位設定モジュールが行うそれぞれのクラスタの順位を再構成する処理は、それぞれのクラスタに含まれるそれぞれの文書がそれぞれの文書対応に付されたスコアの大きい順に並べられている場合、その中央または中央付近に位置するスコアをそれぞれのクラスタごとに得て、そのクラスタごとの中央または中央付近に位置するスコアをそれぞれのクラスタスコアとし、そのクラスタスコアによって、クラスタの順位を再構成することを特徴とする請求項８記載の情報分類装置。

【請求項１２】　前記クラスタリング処理を複数の検索サービスによって得られた検索結果に対応して行うことを可能とする場合、前記クラスタの順位を再構成するためのクラスタスコアを求める処理は、複数の検索サービスそれぞれに対応して行うことを特徴とする請求項９から１１のいずれか１項に記載の情報分類装置。

【請求項１３】　前記クラスタリングモジュールが行うクラスタリング処理は、それぞれの文書のタイトルを検出し、そのタイトルに含まれる特徴的な用語を特徴要素として抽出し、抽出された特徴要素に基づいて行うことを特徴とする請求項８から１２のいずれか１項に記載の情報分類装置。

【請求項１４】　前記クラスタ順位が再構成されたクラスタリング結果の出力の仕方は、クラスタスコアの高いクラスタ順に表示し、クラスタスコアが同じであるクラスタが存在する場合には、クラスタ内の文書数の多いクラスタを高順位とすることを特徴とする請求項８から１３のいずれか１項に記載の情報分類装置。

【請求項１５】　検索サービスにユーザからの検索要求が与えられることによって検索された複数の文書をクラスタリング処理してそのクラスタリング結果を出力する情報分類処理プログラムを記録した記録媒体であって、その情報分類処理プログラムは、

検索サービスによって検索された複数の文書に対し、それぞれの文書の共通性に基づいてクラスタリング処理する手順と、

これによって得られたクラスタリング結果に対し、検索されたそれぞれの文書対応に付された検索要求との適合性を示す値（スコアという）を用いて、前記クラスタリング処理によって得られたそれぞれのクラスタの順位を再構成し、そのクラスタ順位が再構成されたクラスタリング結果を出力する手順と、

を含むことを特徴とする情報分類処理プログラムを記録した記録媒体。

【請求項１６】　前記それぞれのクラスタの順位を再構成する処理は、それぞれのクラスタに含まれるそれぞれの文書対応に付されたスコアの平均値をそれぞれのクラスタごとに求め、クラスタごとの平均値をそれぞれのクラスタのクラスタスコアとし、そのクラスタスコアによって、クラスタの順位を再構成することを特徴とする請求項１５記載の情報分類処理プログラムを記録した記録媒体。

【請求項１７】　前記それぞれのクラスタの順位を再構成する処理は、それぞれのクラスタに含まれるそれぞれの文書対応に付されたスコアの最大値をそれぞれのクラスタごとに得て、そのクラスタごとのスコアの最大値をそれぞれのクラスタスコアとし、そのクラスタスコアによって、クラスタの順位を再構成することを特徴とする請求項１５記載の情報分類処理プログラムを記録した記録媒体。

【請求項１８】　前記それぞれのクラスタの順位を再構成する処理は、それぞれのクラスタに含まれるそれぞれの文書がそれぞれの文書対応に付されたスコアの大きい順に並べられている場合、その中央または中央付近に位置するスコアをそれぞれのクラスタごとに得て、そのクラスタごとの中央または中央付近に位置するスコアをそれぞれのクラスタスコアとし、そのクラスタスコアによって、クラスタの順位を再構成することを特徴とする請求項１５記載の情報分類処理プログラムを記録した記録媒体。

【請求項１９】　前記クラスタリング処理を複数の検索サービスによって得られた検索結果に対応して行うことを可能とする場合、前記クラスタの順位を再構成するためのクラスタスコアを求める処理は、複数の検索サービスそれぞれに対応して行うことを特徴とする請求項１６から１８のいずれか１項に記載の情報分類処理プログラムを記録した記録媒体。

【請求項２０】　前記クラスタリング処理は、それぞれの文書のタイトルを検出し、そのタイトルに含まれる特徴的な用語を特徴要素として抽出し、抽出された特徴要素に基づいて行うことを特徴とする請求項１５から１９のいずれか１項に記載の情報分類処理プログラムを記録した記録媒体。

【請求項２１】　前記クラスタ順位が再構成されたクラスタリング結果の出力の仕方は、クラスタスコアの高いクラスタ順に表示し、クラスタスコアが同じであるクラスタが存在する場合には、クラスタ内の文書数の多いクラスタを高順位とすることを特徴とする請求項１５から２０のいずれか１項に記載の情報分類処理プログラムを記録した記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は汎用の検索サービスで検索された結果に対しクラスタリング処理を施し、かつ、クラスタリング処理によって得られたクラスタの順位を再構成して表示することで、ユーザの検索要求に沿ったクラスタリング結果を提示できるようにした情報分類方法及び情報分類装置並びに情報分類処理プログラムを記録した記録媒体に関する。

【０００２】

【従来の技術】

ネットワーク上に存在する膨大な量の情報の中からユーザの所望とする情報を検索する場合、検索サービスの存在は重要である。たとえば、インターネットでwebページを検索する際、ユーザは、幾つかの検索サービスの中から任意の検索サービスを選び、自分の欲しい情報を得るための検索要求としてのキーワードを入力する。これによって、検索サービス側では、入力されたキーワードに基づいて情報検索を行って、その検索結果をユーザに提示する。

【０００３】

しかし、検索サービスによって検索される情報は膨大な量となることも多く、その中からユーザの本当に欲しい情報を見つけるのは非常に大変である。近年、webページは増大の一途をたどっているため、検索された多数の情報を如何にユ

ーザにわかりやすく提示するかが大きな課題となっている。

【０００４】

最近では、検索された情報をユーザが見やすい形に加工して提示する手法も開発され実用化されつつある。たとえば、ユーザの入力したキーワードで検索された結果から得られるキーワードを用いて再検索することで、情報の絞り込みを行い、ユーザの所望とするwebページを見つけやすくする方法がある。つまり、検索によって得られる検索結果の集合を特徴づけるキーワードを抽出して、ユーザの本当に欲しい情報の集合に収束させる方法である。

【０００５】

このように、膨大な情報の中から、似た情報の集合を見つけることをクラスタリングという。情報処理ではこのクラスタリングはよく知られた手法であり、膨大な文書を分類する場合などに広く使われている。

【０００６】

一般にクラスタリングは、文書の一部、たとえば、webページの場合、文書のタイトル（見出し）などに含まれる特徴的な用語を特徴要素として抽出し、その特徴要素を用いてクラスタリングすることが行われている。勿論、文書全体を解析して特徴要素を抽出し、抽出された特徴要素の基づいてクラスタリングすることも可能であるが、演算量が多く処理に時間がかかる問題がある。

【０００７】

特に、汎用の検索サービスを利用してwebページの検索を行ったのちに、その検索結果の後処理としてクラスタリングを行う場合は、リアルタイムでクラスタリング結果を出力させる必要があり、高速な処理が要求されるので、文書全体を解析してクラスタリングする手法は現実的ではない。その点、上述した文書のタイトルから特徴要素してクラスタリングする手法は、演算量も少なく処理時間も大幅に短縮され、しかも、文書のタイトルは文書の中身を適切に反映したものが多いことから、適切なクラスタリングが可能となるなどの特徴を有している。このため、今後もこの手法は広く使用されるものと考えられる。

【０００８】

【発明が解決しようとする課題】

しかし、上述した文書のタイトルから特徴要素してクラスタリングする手法は、演算量や処理時間の面で優れ、適切なクラスタリングが可能となるとはいっても、クラスタリングを行うための情報量は文書全体から見れば少ないので、全てが適切にクラスタリングされるとは限らない。特に、タイトルが文書の内容を適切に表していなかったり、文書内容とは大きくかけ離れた奇抜なタイトルが付けられていたりする場合もある。このような場合には、クラスタリング精度は大きく低下し、良好なクラスタリング結果は得られないことになる。

【０００９】

また、特徴要素を抽出して、その特徴要素に基づいてクラスタリングする手法は、特徴要素の頻度などを調べ、それによって、機械的に文書を分類してクラスタリングするのが一般的である。このようなクラスタリングでは、文書の意味を解析しているわけではないので、得られたクラスタ（クラスタリングされることによって得られる文書群の１つの集合）が必ずしも意味的な共通性のある文書の集合となるとは限らない。

【００１０】

このように、これまでのクラスタリングは、多数の文書をそのタイトルなどから特徴要素を抽出するなどして、ある程度の適切さを有したクラスタリングがなされるものの、そのクラスタリング結果は、ユーザ側から見て満足度の高いものとはならない場合も多い。

【００１１】

たとえば、ユーザが「半導体」というキーワードをある検索サービスに与えたとき、その検索サービスから得られた多数の文書に対しクラスタリング処理した結果を例にして考える。このクラスタリング結果の一例を図２に示す。

【００１２】

この図２で示されるクラスタリング結果は、クラスタリングによって得られた各クラスタの名称（クラスタ名という）と、それぞれのクラスタに属する検索結果、それぞれのクラスタに幾つの文書が含まれているかを示す文書数などが一覧

表形式で示されている。

【００１３】

なお、検索結果としては、検索された文書のタイトルとそのタイトルに付されたスコアと呼ばれる数値（たとえば、表の最上段にある「東北エプソンの概要」という文書のタイトルに付された「１３３」という数値）が表示される。このスコアは、与えられたキーワードに対しその検索サービスが独自の方法で検索し、検索された文書対応に付けられた値であり、一般には、与えられたキーワードとそれぞれの文書との適合度を示す客観的な尺度として用いられ、汎用の検索サービスによる検索結果にはこのスコアが付されるのが普通である。このスコアは、検索方式により計算の仕方や値の考え方などが異なるものの、一般に、値が大きいほど、与えられたキーワードに適合する内容を有する文書であるといえる。

【００１４】

この図２で示される検索結果において、それぞれのクラスタは文書数の多い順に記述されている。すなわち、「概要」というクラスタ名を持つクラスタ（以下、概要クラスタという）は、そのクラスタ内の文書数が１６個であり、この場合、最も多いので、トップに位置づけられており、続いて、「ＬＰ（レーザプリンタ）」というクラスタ名を持つクラスタ（以下、ＬＰクラスタという）は、そのクラスタ内の文書数が１６個で概要クラスタと同数であるが、ここでは２番目に記述されており、さらに続いて、「仕様」というクラスタ名を持つクラスタ（以下、仕様クラスタという）は、そのクラスタ内の文書数が１４で３番目に記述されており、以下、「デバイス」というクラスタ名を持つクラスタ（以下、デバイスラスタという）は、そのクラスタ内の文書数が９個で４番目、「半導体」というクラスタ名を持つクラスタ（以下、半導体クラスタという）は、そのクラスタ内の文書数が７個で５番目、「電子」というクラスタ名を持つクラスタ（以下、電子クラスタという）は、そのクラスタ内の文書数が４個で６番目に記述されている。

【００１５】

この図２をみると、ユーザの入力したキーワード「半導体」に対して検索された膨大な文書をクラスタリングすると、概要クラスタ、ＬＰクラスタ、仕様クラ

スタ、デバイスクラスタ、半導体クラスタ、電子クラスタなどのクラスタに分類されることがわかる。そして、それぞれのクラスタに含まれる文書数によって図２のようなクラスタ順位での表示がなされる。

【００１６】

ところで、検索結果をクラスタリングしてそのクラスタリング結果を図２のような一覧表形式でユーザ側端末に表示する場合は、ユーザの本当に必要とする情報を一覧表の上位に位置させる方がユーザにとっては見易いものとなるのは言うまでもない。

【００１７】

しかしながら、図２の例では、ユーザの欲する情報が最上位にきているとは思えない。すなわち、概要クラスタ、ＬＰクラスタ、仕様クラスタに含まれる文書の見出しから判断すれば、それらのクラスタには、多種のレーザプリンタに関する仕様概要に関する文書が多く、半導体に直接関係している内容を有する文書は少ないと思われる。この場合、ユーザの与えたキーワードに対してそれに関連しそうな情報が多く含まれそうなクラスタとしては、デバイスクラスタ、半導体クラスタ、電子クラスタであって、これらのクラスタが上位に来ることが望ましいが、この検索結果のクラスタリング結果では、それらのクラスタは下位に位置した状態となっている。

【００１８】

この図２で示されるクラスタリング結果はほんの一例であるが、従来のクラスタリング処理は、膨大な検索結果がクラスタリングされて出力されるものの、そのクラスタリングによって得られるそれぞれのクラスタのクラスタ順位はユーザの与えたキーワードに対し適切な結果とは言えない内容であることがしばしば起こる。

【００１９】

そこで本発明は、汎用の検索サービスで得られた検索結果をクラスタリング処理を施し、かつ、クラスタリングによって得られたクラスタの順位を再構成して表示することで、ユーザの検索要求に沿ったクラスタリング結果を提示できるようにすることを目的としている。

【００２０】

【課題を解決するための手段】

前述の目的を達成するために、本発明の情報分類方法は、検索サービスにユーザからの検索要求が与えられることによって検索された複数の文書に対し、それぞれの文書の共通性に基づいてクラスタリング処理し、それによって得られたクラスタリング結果に対し、検索されたそれぞれの文書対応に付された検索要求との適合性を示す値（スコアという）を用いて、前記クラスタリング処理によって得られたそれぞれのクラスタの順位を再構成し、そのクラスタ順位が再構成されたクラスタリング結果を出力するようにしている。

【００２１】

また、本発明の情報分類装置は、検索サービスにユーザからの検索要求が与えられることによって検索された複数の文書に対し、それぞれの文書の共通性に基づいてクラスタリング処理するクラスタリングモジュールと、このクラスタリングモジュールによって得られたクラスタリング結果に対し、検索されたそれぞれの文書対応に付された検索要求との適合性を示す値（スコアという）を用いて、前記クラスタリング処理によって得られたそれぞれのクラスタの順位を再構成し、そのクラスタ順位が再構成されたクラスタリング結果を出力するクラスタ順位設定モジュールとを有する構成としている。

【００２２】

また、本発明の情報分類処理プログラムを記録した記録媒体は、検索サービスにユーザからの検索要求が与えられることによって検索された複数の文書をクラスタリング処理してそのクラスタリング結果を出力する情報分類処理プログラムを記録した記録媒体であって、その情報分類処理プログラムは、検索サービスによって検索された複数の文書に対し、それぞれの文書の共通性に基づいてクラスタリング処理する手順と、これによって得られたクラスタリング結果に対し、検索されたそれぞれの文書対応に付された検索要求との適合性を示す値（スコアという）を用いて、前記クラスタリング処理によって得られたそれぞれのクラスタの順位を再構成し、そのクラスタ順位が再構成されたクラスタリング結果を出力する手順とを含むようにしている。

【００２３】

これら各発明において、前記それぞれのクラスタの順位を再構成する処理は、それぞれのクラスタに含まれるそれぞれの文書対応に付されたスコアの平均値をそれぞれのクラスタごとに求め、クラスタごとの平均値をそれぞれのクラスタのクラスタスコアとし、そのクラスタスコアによって、クラスタの順位を再構成するようにしている。

【００２４】

このクラスタの順位を再構成する処理は、それぞれのクラスタに含まれるそれぞれの文書対応に付されたスコアの最大値をそれぞれのクラスタごとに得て、そのクラスタごとのスコアの最大値をそれぞれのクラスタスコアとし、そのクラスタスコアによって、クラスタの順位を再構成するようにしてもよく、それぞれのクラスタに含まれるそれぞれの文書がそれぞれの文書対応に付されたスコアの大きい順に並べられている場合、その中央または中央付近に位置するスコアをそれぞれのクラスタごとに得て、そのクラスタごとの中央または中央付近に位置するスコアをそれぞれのクラスタスコアとし、そのクラスタスコアによって、クラスタの順位を再構成するようにしてもよい。

【００２５】

また、前記クラスタリング処理を複数の検索サービスによって得られた検索結果に対応して行うことを可能とする場合、前記クラスタの順位を再構成するためのクラスタスコアを求める処理は、複数の検索サービスそれぞれに対応して行うようにする。

【００２６】

さらに、前記クラスタリング処理は、それぞれの文書のタイトルを検出し、そのタイトルに含まれる特徴的な用語を特徴要素として抽出し、抽出された特徴要素に基づいて行うようにしている。

【００２７】

そして、前記クラスタ順位が再構成されたクラスタリング結果の出力の仕方は、クラスタスコアの高いクラスタ順に表示し、クラスタスコアが同じであるクラスタが存在する場合には、クラスタ内の文書数の多いクラスタを高順位として表

示する。

【００２８】

このように本発明は、クラスタリング結果に対し、検索要求との適合性を示すスコアを用いて、それぞれのクラスタの順位を再構成して出力するようにしているので、ユーザは自分の欲しい情報を探しやすくなる。

【００２９】

そして、それぞれのクラスタの順位を再構成する際に用いられるパラメータとして、クラスタごとのスコアの平均値を用いることで、それぞれのクラスタのキーワードに対する適合性を良好に反映した順位付けを行うことができる。

【００３０】

また、スコアの平均値以外にもクラスタごとのスコアの最大値や中央値を用いることも可能である。最大値を用いることで、クラスタ順位を決めるために足し算や割り算を行う必要がなくなり、計算量を少なくすることができ、しかも、同じクラスタ内に、極端に低いスコアを持つ文書が少数あっても、その影響をあまり受けないようにすることができる。中央値も同様に、クラスタ順位を決めるための計算量を少なくすることができ、しかも、中央値の場合、同じクラスタ内に、極端に高いスコアや極端に低いスコアを持つ文書が少数あっても、その影響をあまり受けないようにすることができる。

【００３１】

また、前記クラスタリング処理は、それぞれの文書のタイトルに含まれる特徴要素に基づいて行うことによって、少ない演算量で適切なクラスタリング処理を行うことができる。

【００３２】

また、複数の検索サービスによって得られた検索結果に対応してクラスタリングを行う場合、クラスタスコアを求める処理を、複数の検索サービスそれぞれに対応して行うようにすれば、特定の検索サービスだけではなく、幾つもの検索サービスで検索された結果に対してもクラスタ順位の再構成処理を行うことができる。

【００３３】

そして、クラスタ順位が再構成されたクラスタリング結果の出力の仕方としては、クラスタスコアの高いクラスタ順に表示し、さらに、クラスタスコアが同じであるクラスタが存在する場合には、クラスタ内の文書数の多いクラスタを高順位とし、それを一覧表形式で表示することにより、ユーザは自分の欲しい情報を探す際、その一覧表を上から順に見て行けばよいことから、効率よく情報をさがすことができる。

【００３４】

【発明の実施の形態】

以下、本発明の実施の形態について説明する。なお、この実施の形態で説明する内容は、本発明の情報分類方法および情報分類装置についての説明であるとともに、本発明の情報分類処理プログラムを記録した記録媒体における情報分類処理プログラムの具体的な処理内容をも含むものである。

【００３５】

図１は本発明を実現するための装置構成を示すもので、大きく分けると、検索サービス１、変換モジュール２、クラスタリングモジュール３、クラスタ順位再構成モジュール４とから構成され、変換モジュール２、クラスタリングモジュール３、クラスタ順位再構成モジュール４が本発明の情報分類装置に相当し、特に、クラスタ順位再構成モジュール４を設けたことに本発明の特徴がある。

【００３６】

検索サービス１はインターネットなどで一般に広く使用されている汎用の検索サービスであり、ユーザからの検索要求としてのキーワードが入力されることにより、たとえばwebページなどから、入力されたキーワードに基づいた情報検索を行う。この検索サービス１で検索された検索結果はファイル形式で出力され、クラスタリングモジュール３に渡されるが、複数の検索サービスが存在する場合、それぞれの検索サービスによって出力されるデータ形式が異なるので、どのような検索サービスからのファイルであってもその内容を読めるような形式にファイルを変換するために変換モジュール２が設けられる。

【００３７】

クラスタリングモジュール３は、検索サービス１により出力された検索結果（変換モジュール２による変換後のファイル内容）に対しクラスタリング処理を行うもので、この実施の形態では、それぞれの文書から文書のタイトルを抽出し、そのタイトルに含まれる特徴的な用語を特徴要素として抽出し、抽出された特徴要素に基づいてクラスタリング処理を行う。

【００３８】

具体的には、それぞれの文書からタイトルとして抽出された部分を形態素解析し、形態素解析された結果から特徴的な用語を特徴要素として抽出する。その後、抽出された特徴要素とそれぞれの文書との関係を示す特徴テーブルを作成する。この特徴テーブルは、たとえば、抽出されたそれぞれの特徴要素が各文書のタイトルに幾つ含まれるているかを、それぞれの特徴要素とそれぞれの文書と対応づけて示すもので、一例として、それぞれの文書のタイトルから、「概要」、「ＬＰ」、「仕様」、「デバイス」、「半導体」、「電子」というような特徴要素が抽出されたとすると、これらの特徴要素が、それぞれの文書のそれぞれのタイトルに、それぞれ何個含まれているかを示す内容となっている。

【００３９】

このような特徴テーブルに基づいて、多数の文書を意味的に共通性のある複数のクラスタに分類する。つまり、それぞれの文書のそれぞれのタイトルに存在する特徴要素に基づいて、共通する特徴要素を持つ文書を１つのまとまりとし、そのまとまりを１つのクラスタとする。

【００４０】

このクラスタリングモジュール３から、たとえば、図２のようなクラスタリング結果が出力されたとする。この図２は前述したように、クラスタリングされることによって得られた各クラスタの名称（上述の特徴要素に相当し、それをここではクラスタ名と呼んでいる）と、そのクラスタに属するそれぞれの文書のタイトルと、そのクラスタには幾つの文書が含まれているかを示す文書数、さらには、前述したように、それぞれのタイトルごとにスコアを示す数値などが一覧表形式で示されている。

【００４１】

なお、このスコアは、前述したように、与えられたキーワードとそれぞれの文書との適合度を示す客観的な尺度として用いられ、ここでは、値が大きいほど、与えられたキーワードに適合する内容を有する文書であるとする。このスコアはキーワードとの適合度を表すものであるので、その単位しては、％や点数など検索サービスによって異なるがこの実施の形態では点数で表すものとする。

【００４２】

そして、このクラスタリングモジュール３によってクラスタリングされた段階のクラスタリング結果は、図２に示されるように、クラスタの配置はそれぞれのクラスタに含まれる文書の数の多い順となっている。前述したように、この場合、上から順に、概要クラスタ、ＬＰクラスタ、仕様クラスタ、デバイスクラスタ、半導体クラスタ、電子クラスタの順となっている。

【００４３】

クラスタ順位再構成モジュール４は、クラスタリングモジュール３で出力されたクラスタリング結果に基づいて、それぞれのクラスタの表示順位を再構成するもので、その処理内容の詳細については後に説明する。

【００４４】

このような構成において、本発明の情報分類処理について説明する。本発明が行う情報分類処理手順は概略的には、図３のフローチャートに示すように、まず、汎用の検索サービス１で検索された検索結果を取得し（ステップＳ１）、取得した検索結果に対しクラスタリング処理を施し（ステップＳ２）、そのクラスタリング結果を出力する（ステップＳ３）。そして、そのクラスタリング結果に対し、それぞれのクラスタ順位の再構成を行い（ステップＳ４）、再構成されたクラスタリング結果を出力する（ステップＳ５）。以下、具体例を参照しながら詳細に説明する。

【００４５】

この実施の形態では、クラスタリングモジュール３が行うクラスタリング処理は、検索サービス１で検索された文書に対し、それぞれの文書のタイトルを抽出し、そのタイトルから特徴要素を抽出して、抽出された特徴要素とそれぞれの文

書との関係を示す特徴テーブルを作成して、その特徴テーブルの内容に基づいて、それぞれの文書を意味的に共通性のある複数のクラスタに分類する。また、この実施の形態では、ユーザが「半導体」というキーワードを検索要求として検索サービス１に入力し、それによって得られた多数の文書がクラスタリングモジュール３によってクラスタリングされ、そのクラスタリング結果が図２に示すような結果であったとする。

【００４６】

このクラスタリングモジュール３からのクラスタリング結果は、クラスタ順位再構成モジュール４に入力され、以下に示すような処理がなされる。

【００４７】

まず、図２で示されるクラスタリング結果における各クラスタ（概要クラスタ、ＬＰクラスタ、仕様クラスタ、デバイスクラスタ、半導体クラスタ、電子クラスタ）において、それぞれのクラスタに含まれる文書対応に付されたスコアを利用して、そのスコアの値の平均を求める。この場合、それぞれのクラスタごとにスコアの値を足し算し、その足し算して得られた結果をそのクラスタに含まれる文書数で割る単純平均を求める。

【００４８】

たとえば、概要クラスタで考えると、この図２に示す検索結果においては、そのクラスタ内のスコアの合計が５７９点あって、文書数が１６個であるので、平均のスコアは約３６点と求められる。また、「ＬＰ」クラスタで考えると、そのクラスタ内のスコアの合計が４５０点であって、文書数が１６個であるので、平均のスコアは約２８点と求められる。同様に、「仕様」クラスタは、そのクラスタ内のスコアの合計が４１３点であって、文書数が１４個であるので、平均のスコアは約２９点と求められ、「デバイス」クラスタは、そのクラスタ内のスコアの合計が８４９点であって、文書数が９個であるので、平均のスコアは約９４点と求められ、「半導体」クラスタは、そのクラスタ内のスコアの合計が７５７点であって、文書数が７個であるので、平均のスコアは約１０８点と求められ、「電子」クラスタは、そのクラスタ内のスコアの合計が３４９点であって、文書数が４個であるので、平均のスコアは約８７点と求められる。

【００４９】

以上のようにして計算された平均のスコアを各クラスタのスコア（クラスタスコアと呼ぶ）とする。そして、このクラスタスコアの高い順にクラスタの順位を再構成する。

【００５０】

すなわち、この場合、クラスタスコアの最も高いクラスタは、半導体クラスタの１０８点であり、第２位はデバイスクラスタの９４点であり、第３位は電子クラスタの８７点であり、以下、概要クラスタ（３６点）、仕様クラスタ（２９点）、ＬＰクラスタ（２８点）といった順序となる。

【００５１】

このようにして、それぞれのクラスタごとにクラスタスコアを計算し、求められたクラスタスコアの高い順にクラスタ順位を再構成する。この再構成されたクラスタリング結果を一覧表形式で表したものが図４である。図４によれば、表の最上段に半導体クラスタが位置し、２番目にデバイスクラスタ、３番目に電子クラスタ、以下、概要クラスタ、仕様クラスタ、ＬＰクラスタといった順序となる。この図４のクラスタリング結果によれば、ユーザの入力した「半導体」というキーワードに対し、そのキーワードに適合する文書が多く含まれるクラスタが上位に来ていることがわかる。

【００５２】

この図４のクラスタリング結果と図２のクラスタリング結果を比較すると、図２のクラスタリング結果では、ユーザの入力した「半導体」というキーワードに対し、そのキーワードとは直接には関係しないような文書で構成される概要クラスタ、ＬＰクラスタ、仕様クラスタといったクラスタが上位に位置し、キーワードに大きく関係するような文書が含まれると思われる半導体クラスタ、デバイスクラスタ、電子クラスタといったクラスタが下位に位置しているが、図４では、それが逆転し、キーワードに大きく関係するような文書が含まれると思われるクラスタが上位に位置するようになる。

【００５３】

なお、クラスタスコアが同じ値となった場合には、クラスタ内に含まれる文書

数の多い方を上位とするなどの措置を講ずる。

【００５４】

以上説明したように、単純にそれぞれのクラスタに含まれる文書数（１つのクラスタにまとめられた文書数）によって順位付けするのではなく、それぞれのクラスタごとにそのクラスタに含まれる文書に付されたスコアに基づいてクラスタの順位を決めることによって、キーワードに適合したクラスタ順位が得られる。

【００５５】

なお、図４に示すようなクラスタリング結果がユーザに表示され、ユーザはこのようなクラスタリング結果の一覧表を見て、自分の欲しい情報の入っていそうな文書のタイトル部分をクリックすれば、そのタイトルに対応する本文が表示されるというような表示処理がなされる。

【００５６】

以上説明したように、この実施の形態では、ユーザの入力したキーワードによって検索された多数の文書に対し、これら多数の文書のタイトルに含まれる特徴要素に基づいてクラスタリング処理し、さらに、そのクラスタリング結果に対して、それぞれのクラスタごとにそのクラスタに含まれる文書のスコアの平均を求める。そして、その平均のスコアをクラスタスコアとし、それぞれのクラスタごとのクラスタスコアに基づいて、クラスタ順位の再構成を行う。つまり、クラスタスコアの大きい順にクラスタの並べ替えを行い、図４に示すようなクラスタリング結果として表示する。

【００５７】

これによって、ユーザの欲しい情報の入っていそうなクラスタが一覧表の上位に位置した状態で表示されているので、自分の欲しい情報を探しやすくなる。

【００５８】

また、これまでの説明では、ある１つの汎用の検索サービスで検索された結果をクラスタリング処理する場合について説明したが、複数の検索サービスにより検索された結果をクラスタリング処理する場合にも適用できる。

【００５９】

検索サービスはそれぞれに得意の分野があることも多く、たとえば、ある検索

サービスはスポーツ関係の情報を多数保有し、ある検索サービスは学術関係の情報を多数保有し、また、ある検索サービスは芸能関係の情報を多数保有しているというように、それぞれの得意の分野が存在する場合も多い。これらそれぞれの得意分野については豊富な情報を所有しており、ユーザの所望とする情報が適切に取り出される可能性が高い。したがって、情報検索を行う際は、検索サービスを使い分けることも普通に行われる。

【００６０】

このように、複数の検索サービスを用いてクラスタリング処理する場合には、それぞれの検索サービスにより検索された検索結果の内容、長さ、検索結果出力順序などがまちまちなので、それぞれの検索サービスからのファイルをクラスタリングモジュール３で処理可能な形式に変換する変換モジュール２を複数の検索サービスに対応して用意する。そして、さらに、そのクラスタリング結果におけるクラスタ順位再構成を行う場合には、それぞれのクラスタのクラスタスコアを求める処理をそれぞれの検索サービスに対応して行うようにする。

【００６１】

たとえば、本発明の要旨であるクラスタ順位再構成処理についていえば、検索サービスによって幾つかの対策を講じる必要がある。たとえば、スコアの分布の幅が非常に大きい場合（たとえば、スコアを表す数値が最大１０００から最小は２など）は、対数を取って計算するなどの措置を講じたり、また、きわめてスコアの値が小さい文書（たとえば、殆どの文書が数百のスコアの値があるのに２や３の値しかない文書）はクラスタリング対象から外すといった措置を講じる。

【００６２】

このように、複数の検索サービスに対応できるようにすることで、ユーザは検索しようとする情報の分野に応じて検索サービスを使い分けることができ、それぞれの得意分野に応じた検索が可能となるばかりでなく、ある１つの検索サービスが混み合っているような場合には、他の検索サービスに切り換えて検索を行うというような柔軟な検索も可能となる。

【００６３】

なお、本発明は以上説明した実施の形態に限定されるものではなく、本発明の

要旨を逸脱しない範囲で種々変形実施可能となるものである。たとえば、これまで説明した実施の形態では、それぞれのクラスタのクラスタスコアは、そのクラスタに含まれる文書のスコアの単純平均を用いた例について説明したが、このクラスタスコアとしては、それぞれのクラスタ内に含まれる文書のなかで最大のスコアを有する文書のスコアを用いるようにしてもよく、また、それぞれのクラスタ内に含まれる文書に付されたスコアのなかで中央に位置する文書のスコアを用いるようにしてもよい。

【００６４】

このように、クラスタごとのスコアの最大値を用いることで、クラスタ順位を決めるために足し算したり割り算したりという計算を行う必要がなく、計算量を少なくすることができ、しかも、同じクラスタ内に、極端に低いスコアを持つ文書が少数あっても、その影響をあまり受けないようにすることができる。また、クラスタごとのスコアの中央値を用いる場合も、最大値を用いるのと同様、クラスタ順位を決めるための計算量を少なくすることができ、しかも、中央値の場合、同じクラスタ内に、極端に高いスコアや極端に低いスコアを持つ文書が少数あっても、その影響をあまり受けないようにすることができる。

【００６５】

また、前述の実施の形態では、クラスタリングを行うための情報（クラスタリング対象情報）として、検索されたそれぞれの文書のタイトルを用いた例について説明したが、これは、タイトルだけでなく、たとえば、ＵＲＬアドレス（http://を取り除いた部分）、更新日時（単純な時間または最近１カ月以内の更新日時）、ファイルサイズ（webページ本文のバイトサイズなど）を用いてクラスタリングすることもできる。また、これらは、単独で用いてクラスタリングするようにしてもよく、幾つかを組み合わせてもよい。このように、クラスタリング対象情報を種々選ぶことによって、それぞれに応じた特色のあるクラスタリングが行える。そして、これらのどれを用いるかは、最初にメニューなどで選択項目を選ぶことで可能となる。また、選んだ項目が無い場合には、他の項目を代用する。たとえば、タイトルを選んだ場合、webページにタイトルが無い場合には、ＵＲＬアドレスを代用する。

【００６６】

また、以上説明した本発明の情報分類処理を行う処理プログラムは、フロッピィディスク、光ディスク、ハードディスクなどの記録媒体に記録させておくことができ、本発明はその記録媒体をも含むものである。また、ネットワークから処理プログラムを得るようにしてもよい。

【００６７】

【発明の効果】

以上説明したように本発明によれば、検索された複数の文書をクラスタリング処理し、そのクラスタリング処理結果に対し、検索要求との適合性を示すスコアを用いて、それぞれのクラスタの順位を再構成して出力し、クラスタスコアの高い順に表形式で表示するようにしているので、ユーザは自分の欲しい情報を探す際、その一覧表を上から順に見て行けばよいことから、効率よく情報をさがすことができる。しかも、クラスタ順位再構成処理としては、もともと存在するスコアを用いて、それぞれのクラスタごとに平均値や最大値あるいは中央値などを求め、その結果に基づいて並べ替えするだけの処理であるので、複雑な演算処理が不要で短時間での処理が可能となる。これによって、汎用の検索サービスを利用してwebページの検索を行ったのちに、その検索結果の後処理として本発明を適用する場合にも、殆どリアルタイムで検索結果（クラスタ順位再構成後の最終的なクラスタリング結果）をユーザに提示することが可能となる。

【図面の簡単な説明】

【図１】

本発明の情報分類装置の実施の形態を説明する構成図である。

【図２】

ある検索サービスで検索された検索結果としての複数の文書をクラスタリングした結果の一例を示す図である。

【図３】

本発明の情報分類処理手順を概略的に説明するフローチャートである。

【図４】

図２で示されたクラスタリング結果をクラスタ順位再構成処理した結果を示す

図である。

【符号の説明】

1　検索サービス

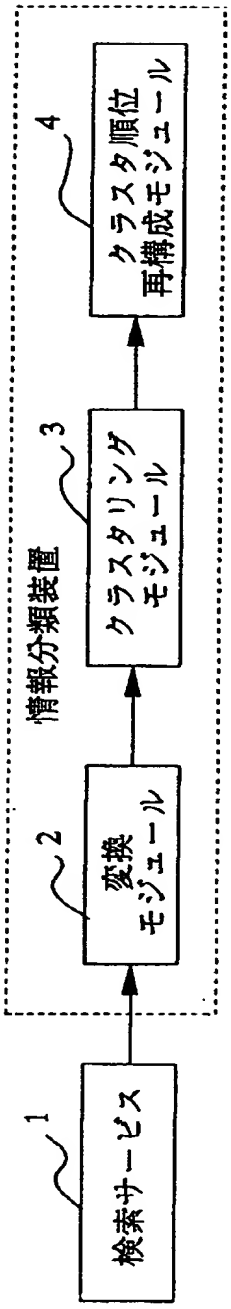
2　変換モジュール

3　クラスタリングモジュール

4　クラスタ順位再構成モジュール

【書類名】　図面

【図１】

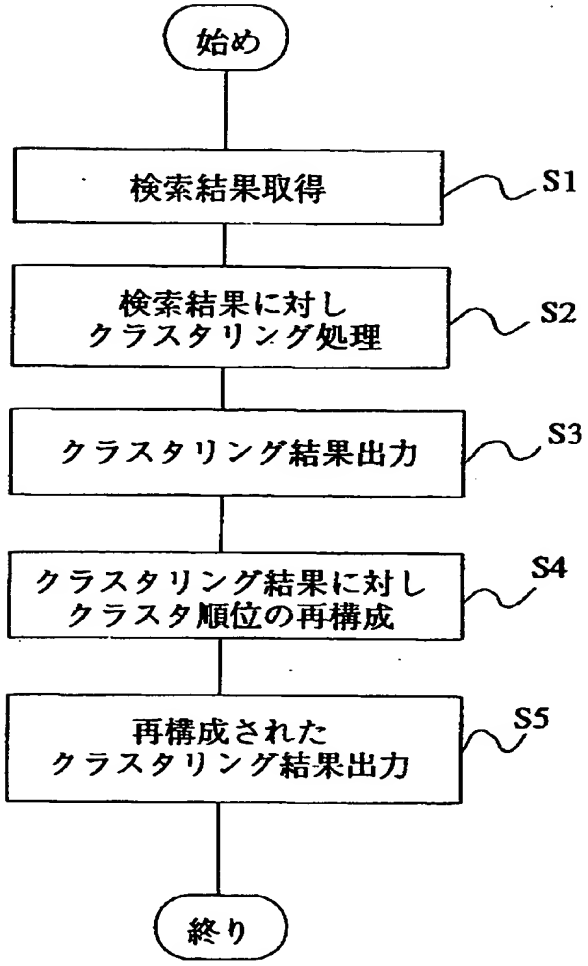


【図２】

スコア　　文書のタイトル

クラスタ名	文書数	検索結果
概要	16	133　東北エプソンの概要 49　LP-7000 仕様概要 38　LP-8000 仕様概要 38　LP-8400 仕様概要 34　LP-700 仕様概要 33　セイコーエプソン会社概要 30　LP-9200 仕様概要 28　LP-9200PS2 仕様概要 28　LP-830PS 仕様概要 27　LP-8300 仕様概要 26　LP-8600 仕様概要 24　LP-8200 仕様概要 23　LP-710 仕様概要 23　LP-800 仕様概要 23　LP-900 仕様概要 22　LP-500 仕様概要 （クラスタ内合計　579　平均　36）
LP	16	49　LP-7000 仕様概要 38　LP-8000 仕様概要 38　LP-8400 仕様概要 34　LP-700 仕様概要 30　LP-700S Spec Sheet & Option 30　LP-9200 仕様概要 28　LP-9200PS2 仕様概要 28　LP-830PS 仕様概要 27　LP-8300 仕様概要 26　LP-8600 仕様概要 24　LP-8200 仕様概要 23　LP-710 仕様概要 23　LP-800 仕様概要 23　LP-900 仕様概要 22　LP-500 仕様概要 7　エスパーレーザー　LP-8300 （クラスタ内合計　450　平均　28）
仕様	14	49　LP-7000 仕様概要 38　LP-8000 仕様概要 38　LP-8400 仕様概要 34　LP-700 仕様概要 30　LP-9200 仕様概要 28　LP-9200PS2 仕様概要 28　LP-830PS 仕様概要 27　LP-8300 仕様概要 26　LP-8600 仕様概要 24　LP-8200 仕様概要 23　LP-710 仕様概要 23　LP-800 仕様概要 23　LP-900 仕様概要 22　LP-500 仕様概要 （クラスタ内合計　413　平均　29）
デバイス	9	117　デバイス　半導体　CARD-PC 115　デバイス　半導体　メモリ 111　電子デバイス　ＡＳＩＣ 111　デバイス　半導体　ＰＣカード製品 101　デバイス　コラボレーション 90　電子デバイス　半導体　マイコン 82　ＥＰＳＯＮ　電子デバイス新製品 66　ＥＰＳＯＮ　電子デバイスお問い合せ 56　デバイス　半導体　ＡＳＳＰ （クラスタ内合計　849　平均　94）
半導体	7	157　半導体事業部現場方針 117　デバイス　半導体　CARD-PC 115　デバイス　半導体　メモリ 111　電子デバイス　ＡＳＩＣ 111　デバイス　半導体　ＰＣカード製品 90　電子デバイス　半導体　マイコン 56　デバイス　半導体　ＡＳＳＰ （クラスタ内合計　757　平均　108）
電子	4	111　電子デバイス　ＡＳＩＣ 90　電子デバイス　半導体　マイコン 82　ＥＰＳＯＮ　電子デバイス新製品 66　ＥＰＳＯＮ　電子デバイスお問い合せ （クラスタ内合計　349　平均　87）

【図３】


始め
検索結果取得
S1
検索結果に対し
クラスタリング処理
S2
クラスタリング結果出力
S3
クラスタリング結果に対し
クラスタ順位の再構成
S4
再構成された
クラスタリング結果出力
S5
終り

【図４】

スコア　　文書のタイトル

クラスタ名	文書数	検索結果
半導体	7	157 半導体事業部環境方針 117 デバイス 半導体 CARD-PC 115 デバイス 半導体 メモリ 111 電子デバイス ＡＳＩＣ 111 デバイス 半導体 ＰＣカード製品 90 電子デバイス 半導体 マイコン 56 デバイス 半導体 ＡＳＳＰ （クラスタ内合計 757 平均 108）
デバイス	9	117 デバイス 半導体 CARD-PC 115 デバイス 半導体 メモリ 111 電子デバイス ＡＳＩＣ 111 デバイス 半導体 ＰＣカード製品 101 デバイス コラボレーション 90 電子デバイス 半導体 マイコン 82 ＥＰＳＯＮ 電子デバイス新製品 66 ＥＰＳＯＮ 電子デバイスお問い合せ 56 デバイス 半導体 ＡＳＳＰ （クラスタ内合計 849 平均 94）
電子	4	111 電子デバイス ＡＳＩＣ 90 電子デバイス 半導体 マイコン 82 ＥＰＳＯＮ 電子デバイス新製品 66 ＥＰＳＯＮ 電子デバイスお問い合せ （クラスタ内合計 349 平均 87）
概要	16	133 東北エプソンの概要 49 LP-7000 仕様概要 38 LP-8000 仕様概要 38 LP-8400 仕様概要 34 LP-700 仕様概要 33 セイコーエプソン会社概要 30 LP-9200 仕様概要 28 LP-9200PS2 仕様概要 28 LP-830PS 仕様概要 27 LP-8300 仕様概要 26 LP-8600 仕様概要 24 LP-8200 仕様概要 23 LP-710 仕様概要 23 LP-800 仕様概要 23 LP-900 仕様概要 22 LP-500 仕様概要 （クラスタ内合計 579 平均 36）
仕様	14	49 LP-7000 仕様概要 38 LP-8000 仕様概要 38 LP-8400 仕様概要 34 LP-700 仕様概要 30 LP-9200 仕様概要 28 LP-9200PS2 仕様概要 28 LP-830PS 仕様概要 27 LP-8300 仕様概要 26 LP-8600 仕様概要 24 LP-8200 仕様概要 23 LP-710 仕様概要 23 LP-800 仕様概要 23 LP-900 仕様概要 22 LP-500 仕様概要 （クラスタ内合計 413 平均 29）
LP	16	49 LP-7000 仕様概要 38 LP-8000 仕様概要 38 LP-8400 仕様概要 34 LP-700 仕様概要 30 LP-700S Spec Sheet & Option 30 LP-9200 仕様概要 28 LP-9200PS2 仕様概要 28 LP-830PS 仕様概要 27 LP-8300 仕様概要 26 LP-8600 仕様概要 24 LP-8200 仕様概要 23 LP-710 仕様概要 23 LP-800 仕様概要 23 LP-900 仕様概要 22 LP-500 仕様概要 7 エスパーレーザー LP-8300 （クラスタ内合計 450 平均 28）

【要約】

【課題】汎用の検索サービスによって検索された多数の文書をクラスタリングし、それを一覧表形式で表示する際、そのクラスタリングによって得られたクラスタの並びが、ユーザの与えたキーワードにあまり関係のない文書を含むクラスタが上位に位置してしまう場合がある。

【解決手段】汎用の検索サービス１で検索された複数の検索結果をクラスタリング処理し、さらに、そのクラスタリング結果に対し、クラスタ順位再構成ジュール４がそれぞれのクラスタに含まれるそれぞれの文書対応に付された前記検索要求との適合性を示すスコアを用い、クラスタごとにスコアの平均を求めるなどしてクラスタごとのスコア（クラスタスコア）を求め、それによってクラスタの順位を再構成し、クラスタスコアの高い順に表形式で出力する。

【選択図】　図１

出証特２０００－３０５２０４３

出　願　人　履　歴　情　報

識別番号　　〔０００００２３６９〕

１．変更年月日　　１９９０年　８月２０日
［変更理由］　新規登録
住　所　東京都新宿区西新宿２丁目４番１号
氏　名　セイコーエプソン株式会社